Task-specific Word-Clustering for Part-of-Speech Tagging

Yoav Goldberg
Google, Inc.
`yogo@google.com`
`yoav.goldberg@gmail.com`

Abstract

While the use of cluster features became ubiquitous in core NLP tasks, most cluster features in NLP are based on distributional similarity. We propose a new type of clustering criteria, specific to the task of part-of-speech tagging. Instead of distributional similarity, these clusters are based on the behavior of a baseline tagger when applied to a large corpus. These cluster features provide similar gains in accuracy to those achieved by distributional-similarity derived clusters. Using both types of cluster features together further improve tagging accuracies. We show that the method is effective for both the in-domain and out-of-domain scenarios for English, and for French, German and Italian. The effect is larger for out-of-domain text.

1 Introduction

The limited amounts of annotated training data available for supervised learning call for semi-supervised learning approaches, which aim to leverage the vast amounts of readily available unannotated data in order to improve the accuracies of supervised systems.

In natural-language processing, a simple and popular method for semi-supervised learning is based on word clustering (Miller et al., 2004; Koo et al., 2008; Turian et al., 2010): words in a large corpus are clustered into equivalence classes based on some (usually distributional) criteria, and the induced classes are then used as additional features in a supervised learning model. The use of cluster-based features was demonstrated to improve the accuracies of many NLP tasks, including parsing (Koo et al., 2008; Candito and Crabbé, 2009), named-entity recognition (Miller et al., 2004; Turian et al., 2010; Lin and Wu, 2009; Chrupala, 2011), classification of semantic relations (Chrupala, 2011) and machine-translation (Uszkoreit and Brants, 2008).

Word clusters are usually induced based on a distributional-similarity criteria: words are clustered based on the words that tend to occur before or after them. Clusters produced by the Brown clustering algorithm (Brown et al., 1992) are an example of commonly used distributional clustering features. In this model, words are clustered by means of a probabilistic cluster-based language model. A more scalable distributional clustering algorithm is introduced by Uszkoreit and Brants (2008) who use a parallel implementation of the Exchange algorithm to cluster words based on word-to-cluster transitions. When used as features, clusters derived using the Exchange algorithm are as effective as those derived by the Brown algorithm. Other types of distributional clustering algorithms rely on word embeddings (Collobert and Weston, 2008; Mnih and Hinton, 2007). Turian et al. (2010) find that embedding-based distributional clusters tend to underperform Brown-type clusters.

The distributional-similarity hypothesis underlying all these algorithms is that words in similar contexts behave in a similar manner. The notion of similarity is vague and is not specific to any particular task. In practice, distributional-similarity based clusters show a mix of semantic and syntactic properties. Can we design word-clusters capturing properties which are relevant for a specific task?

We focus on the task of part-of-speech (POS) tagging, and present a novel *task-specific* clustering criteria: words are clustered based on the behavior of a baseline tagger when applied to a large body of text.

One of the most useful sources of information for tagging a given word w with a tag t is the weighted ambiguity class of the word, as represented by the conditional tagging distribution $p(T = t|w)$. Our first kind of clusters aim to capture exactly this information: we cluster words based on their empirical $p(T = t|w)$ distributions, as observed over a large automatically tagged corpus.

The tag of a word w can be predicted to some extent also by the previous word w_{-1} or the following one w_{+1}. We can create word-clusters to capture these sources of information by clustering words based on the empirical distributions $p(T_{+1} = t|w)$ and $p(T_{-1} = t|w)$, and using these clusters to represent the previous and following word respectively.

The approach is related to self-training in that we use the tagger's own prediction in order to improve it. However, in contrast to self training, we use statistics *derived from* the tagger's output as *additional features* for supervised training.

For POS-tagging, our task-specific clusters are as effective as those derived by a lexical distributional-similarity criteria when used on their own, and have a cumulative effect when both kinds of clusters are used together. Moreover, the task-specific clusters serve as a very good proxy to word identity for the purpose of POS-tagging, and we can train completely unlexicalized POS-tagging models without sacrificing accuracy.

2 Method

Our training protocol is as follows:
1) Train a supervised tagger on POS-annotated text.
2) Use the tagger to annotate large amounts of raw text.
3) Collect (W, T) counts from the automatically tagged text.
4) For a word w occurring over k times, compute:
 $p(T = t|w) = count(w, t)/\sum_{t'} count(w, t')$
5) Cluster words based on $p(T = t|w)$.

We then use the derived clusters as additional features in a discriminatively-trained sequence-tagger.

When clustering, we encode the conditional $p(T = t|w)$ distribution for each word w as a $|T|$-dimensional vector in which the ith entry is the conditional probability $p(T = t_i|w)$, and cluster words based on the Euclidean distance between their vectors:

$$dist(w_1, w_2) = \sqrt{\sum_{t_i \in T} (p(t_i|w_1) - p(t_i|w_2))^2}$$

We use the K-means clustering algorithm with the initialization procedure described in (Arthur and Vassilvitskii, 2007), which stochastically favors cluster centers that are far apart from previously chosen centers.

Words provide weak signals regarding the POS-tag of the next or previous word. We produce clusters based on the distributions $p(T_{-1} = t|w)$ and $p(T_{+1} = t|w)$ in a similar fashion.

3 Details and Experiments

3.1 Parameters

In all the experiments, we set the word frequency threshold k to be 100. Due to the large size of our unannotated corpus, we still remain with very large vocabulary sizes (see Table 2). We run the K-means algorithm for 100 iterations, and cluster the words into 256 classes. While the baseline-tagger features are tuned for good accuracy, we did not perform all but minimal tuning of the extended cluster features, and did not tune any of the other parameters.

3.2 Tagger

We use a first-order linear-chain sequence-tagger[1], trained using the averaged structured-MIRA algorithm. The features include distributional clusters derived from the unannotated corpora using the Exchange algorithm and are detailed in Table 1. Throughout the presentation, all features are assumed to be conjoined with the tag to be predicted.

---

[1]Most previous work on POS-tagging, e.g. (Ratnaparkhi, 1996; Brants, 2000; Collins, 2002; Toutanova et al., 2003) use at-least a second-order model for their better results. In contrast, we use a first-order model which is much faster. Thus, our tagging results are lower than reported in previous work evaluating on the WSJ corpus (our train/test split is also somewhat different). Our primary interest in this work is not in demonstrating state-of-the-art tagging accuracies on the WSJ corpus but rather examining the contributions of different cluster features to the tagger accuracy on diverse corpora.

Type	Templates
Lexical	w_0
Signature	$pref(1)\ pref(2)\ pref(3)$
	$suf(1)\ suf(2)\ suf(3)$
	$capitalization\ hyphen$
Transition	t_{-1}
ρ Cluster-Dist	$\rho_0 \quad \rho_{-1} \quad \rho_{-2} \quad \langle \rho_{-1}, \rho \rangle \quad \langle \rho_{-2}, \rho_{-1} \rangle$
+Transition	$\langle \rho_0, t_{-1} \rangle \quad \langle \rho_{-1}, t_{-1} \rangle$

Table 1: Tagger features for the baseline tagger. pref(n) and suf(n) are prefixes/suffixes of length n of the current word w_0. The distributional-similarity features ρ are derived using the algorithm of (Uszkoreit and Brants, 2008).

3.3 Datasets

Annotated data For English, we use the following annotated corpora:

WSJ The WSJ portion of the Penn Treebank corpus (Marcus et al., 1993) is used to train all of our English tagging models. We train on Sections 2-21, and evaluate on Section 22.

Brown (BRN) The entire Brown corpus portion of the Penn Treebank is used for evaluation.

Questions (QTB) The QuestionBank (Judge et al., 2006) contain 4,000 questions, which we use for evaluation,

Football (FTBL) We report results on the development set (185 sentences) of the Football corpus of (Foster, 2010). In one experiment we use the test section (170 sentences) as additional training data.

Web The entire *web* portion of the Ontonotes corpus (Weischedel et al., 2011) is used for evaluation.[2]

In most experiments we train our tagger on the training set of the WSJ corpus and reserve the other datasets for evaluation. The baseline tagger is always trained on the WSJ training set.

German We use data and splits from the CoNLL 2006 shared task (Buchholz and Marsi, 2006).

French We use the French Treebank (Abeillé and Barrier, 2004) with splits defined in Candito et al. (2010).

---

[2]Note that the Ontonotes corpus is systematically different from the training corpus in several aspects, including using both the "IN" and "TO" tags for the word "to" depending on its usage (in the Penn Treebank all, *to* is consistently tagged as TO), and the introduction of additional tags for hyphens and non-sentence-final punctuation. While one could get vastly improved accuracies on this dataset by specifically addressing these issues, we did not do so in the current work as our primary interest is comparing the effect of the various cluster features on tagging accuracy.

Language	Domain	#Tokens	Vocabulary
English	News	19×10^9	649K
German	News	2.5×10^9	386K
French	News	1.4×10^9	165K
Italian	News	0.5×10^9	116K

Table 2: Details of unannotated data. Vocabulary is the number of token-types appearing more than 100 times.

Italian We use data and splits from the CoNLL 2007 shared task (Nivre et al., 2007).

Unannotated Data We use one year of newswire articles from multiple sources from a news aggregation website for each language. The datasets range in size from 19 to 0.5 billion tokens. The unannotated data is summarized in Table 2.

4 Results

4.1 English

In the first set of experiments we test the effectiveness of the Task-based clustering method on both in-domain and out-of-domain English data.

We begin by distilling the amount of information captured by the different clusters. To this end, we train models with the simplest set of features possible: for each sequence position we consider the lexical item w_0, the transition feature t_{-1}, and zero or more cluster features. We also train models including the cluster features but not the lexical items. We evaluate the models on the different English datasets. Table 3 detail the results.

Features	WSJ	QTB	BRN	FTBL	Web
$t_{-1}\ w_0$	94.97	85.93	91.29	89.79	88.77
$t_{-1}\ w_0\ \rho_0$	96.01	88.28	94.11	91.69	91.13
$t_{-1}\ w_0\ \zeta_0$	**96.42**	**89.74**	**94.95**	92.85	**92.17**
$t_{-1}\ w_0\ \eta_{-1}$	95.21	86.07	91.49	89.58	89.10
$t_{-1}\ w_0\ \tau_{+1}$	95.30	86.34	91.64	89.73	89.04
$t_{-1}\ \rho_0$	94.37	85.66	92.17	89.49	88.97
$t_{-1}\ \zeta_0$	96.14	89.24	94.74	**93.17**	92.06

Table 3: Tagging accuracies with minimal feature sets. ρ: distributional-clusters, ζ: $p(t|w)$-clusters, η: $p(t_{+1}|w)$-clusters, τ: $p(t_{-1}|w)$-clusters. All models are trained on the training portion of the WSJ corpus.

Note that this is not exactly a domain-adaptation scenario, as all the unannotated data is from the Newswire domain. Still, the cluster features contribute to tagging accuracies across all the datasets.

When the current word w_0 is present as feature, the distributional clusters ρ_0 is somewhat less informative than the task-specific clustering ζ_0, which is based on $p(t|w)$. The cluster features of the next and previous words (η_{-1} and τ_{+1}) are expectedly less informative than the cluster associated with the current word, but still contain some predictive information. When we exclude the word from the feature set and rely only on the cluster information (the last two rows of the table), the task-specific clusters ζ_0 do particularly well – compensating almost completely over the missing word identity information. The models relying solely on the previous tag and the task-specific cluster ($t_{-1}\ \zeta_0$) are significantly better than the models relying on the previous tag and the explicit word identity ($t_{-1}\ w_0$).

We then proceed to evaluate the effectiveness of the cluster features in the context of a richer feature set. We use the feature-sets described in Tables 1 and 4. We consider different subsets of the cluster features. Results are presented in Table 5.

Type	Templates				
ζ Cluster $p(t|w)$ +Transition	ζ_0	ζ_{-1}	ζ_{-2}	$\langle\zeta_{-1},\zeta_0\rangle$	$\langle\zeta_{-2},\zeta_{-1}\rangle$
		$\langle\zeta_0,t_{-1}\rangle$		$\langle\zeta_{-1},t_{-1}\rangle$	
η Cluster $p(t_{+1}|w)$ +Transition	η_{-1}				
	$\langle\eta_{-1},t_{-1}\rangle$				
τ Cluster $p(t_{-1}|w)$	τ_{+1}				

Table 4: Additional cluster features.

	No clusters	Dist ρ	Task ζ	Dist+Task $\rho\,\zeta$	All $\rho\,\zeta\,\eta\,\tau$	All (no w_0)
WSJ	96.35	96.90	96.82	**97.01**	97.02	97.02
QTB	88.86	90.74	90.50	90.83	**90.93**	90.93
BRN	94.37	95.57	95.48	95.68	**95.72**	95.70
FTBL	91.96	93.38	93.44	93.74	93.80	**94.03**
Web	91.38	92.81	92.82	92.99	**93.05**	93.05

Table 5: Tagging accuracies using the different clusterings within a rich feature set. All models are trained on the training portion of the WSJ corpus. Last column contain all the features but the lexical one.

Expectedly, using the richer feature-sets improve results for all models. The cluster features still contribute to tagging accuracies across all the datasets. The contribution of the task-based clusters (Task) is similar but a bit lower than that of the distributional clusters (Dist), but results improve when the two clustering approaches are combined (Dist+Task). Adding the task based clustering of neighboring

words (All) further improve the results on most datasets. The largest improvements are observed on the out-of-domain datasets. Somewhat surprisingly, dropping the explicit lexical feature w_0 (last column) does not hurt performance, and even significantly improve it on the Football dataset.

4.2 English – Additional Training Data

In the next experiment, we target the situation in which we have a small amount of annotated data in an interest-domain in addition to the larger amount of out-of-domain data. We use the test-set of the Football dataset as additional in-domain training material. Results are presented in Table 6. As expected, using the additional in-domain training data improve the results. However, the contribution of the additional data is small, as most of the gap is already covered by the cluster features. When using all the cluster features but no lexicalization (last column) training on WSJ alone outperform the joint training.

Train	No clusters	Dist ρ	Task ζ	Dist+Task $\rho\,\zeta$	All $\rho\,\zeta\,\eta\,\tau$	All (no w_0)
WSJ	91.96	93.38	93.44	93.74	93.80	**94.03**
+FTBL	92.31	93.47	93.50	93.92	**93.94**	93.83

Table 6: Adaptation results to the Football domain when training on both datasets.

4.3 German, French and Italian

We observe similar trends on languages other than English (Table 7). The additional task-specific cluster features improve performance across all languages.

Language	No Clusters	Dist ρ	Task ζ	Dist+Task $\rho\,\zeta$	All $\rho\,\zeta\,\eta\,\tau$
German	96.48	97.68	97.70	97.84	**98.00**
French	96.45	97.55	97.54	97.66	**97.74**
Italian	93.58	96.00	96.15	**96.43**	96.39

Table 7: Tagging results for German, French and Italian.

5 Conclusions

We presented a task-specific word clustering method for POS-tagging. The method is effective across domains and languages. The automatically derived clusters capture the essence of the lexical items with respect to the task to the extent that the cluster features can replace the actual lexical items. We would like to see task-specific clusterings for other, more challenging tasks.

References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a french treebank. In *Proceedings of LREC*.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *SODA*.

T. Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proc. of ANLP*.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*.

Marie Candito and Benoît Crabbé. 2009. Improving generative statistical parsing with semi-supervised word clustering. In *Proc. of IWPT*.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of LREC*.

Grzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with lda. In *Proc. of IJCNLP*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proc. of ACL*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML*.

Jennifer Foster. 2010. "cba to check the spelling" investigating parser performance on discussion forum posts. In *Proc. of HLT-NAACL*.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proc. of ACL*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-HLT*.

Dekang Lin and Xiaoyun Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of ACL-IJCNLP*, pages 1030–1038.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT-NAACL*.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of ICML*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL*.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of EMNLP*.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of HLT-NAACL*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of ACL-HLT*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. OntoNotes release 4.0. Linguistic Data Consortium, Philadelphia.